First occurrence in pairs of long words: a Penney-ante conjecture of Pevzner

Stark, D

Abstract: Suppose X1,X2, is a sequence of independent and identically distributed random elements taking values in a finite set S of size |S| 2 with probability distribution P(X=s)=p(s)>0 for s S. P. Pevzner [Kvantl 5 (1987), 4–15; per bibl.] has conjectured that for every probability distribution P there exists an N>0 such that for every word A with letters in S whose length is at least N, there exists a second word B of the same length as A, such that the event that B appears before A in the sequence X1,X2, has greater probability than that of A appearing before B. In this paper it is shown that a distribution P satisfies Pevzner's conclusion if and only if the maximum value of P, p, and the secondary maximum c satisfy the inequality c>p(1−p)/(1+p). For |S|=2 or |S|=3, the inequality is true and the conjecture holds. If c p(1−p)/(1+p), then the conjecture is true when A is not allowed to consist of pure repetitions of that unique element for which the distribution takes on its mode.

DOI: https://doi.org/10.1017/S0963548300001656

Combinatorics, Probability and Computing (1995) 4, 279–285

First Occurrence in Pairs of Long Words: A Penney-ante Conjecture of Pevzner

DUDLEY STARK†

Institut für Angewandte Mathematik, Universität Zürich,

Winterthurerstr. 190, CH-8057 Zürich, Switzerland

Received 10 September 1993; revised 8 August 1994

Suppose $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random elements whose values are taken in a finite set S of size $|S| \geq 2$ with probability distribution $\mathbb{P}(X = s) = p(s) > 0$ for $s \in S$. Pevzner has conjectured that for every probability distribution $\mathbb{P}$ there exists an $N > 0$ such that for every word A with letters in S whose length is at least N, there exists a second word B of the same length as A, such that the event that B appears before A in the sequence $X_1, X_2, \ldots$ has greater probability than that of A appearing before B. In this paper it is shown that a distribution $\mathbb{P}$ satisfies Pevzner's conclusion if and only if the maximum value of $\mathbb{P}$, p, and the secondary maximum c satisfy the inequality $c > p\frac{1-p}{1+p}$. For $|S| = 2$ or $|S| = 3$, the inequality is true and the conjecture holds. If $c \leq p\frac{1-p}{1+p}$, then the conjecture is true when A is not allowed to consist of pure repetitions of that unique element for which the distribution takes on its mode.

1. Introduction

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random elements taking on values in some finite space S of size $|S| \geq 2$ with distribution $\mathbb{P}(X_i = s) = p(s) > 0$. We will call an element of S a *letter* and a finite sequence of elements of S a *word*. Given words $A = a_1 a_2 \ldots a_n$ and $B = b_1 b_2 \ldots b_\eta$ of length n and η, respectively, we define

$$T_A \equiv \inf\{k : X_{k-n+1} X_{k-n+2} \ldots X_k = a_1 a_2 \ldots a_n\}$$

to be the time of first occurrence of A in the sequence X_i, with a similar definition for T_B.

For words A and B, let $Bwins = \{T_B < T_A\}$ be the event that B appears in the sequence $X_1, X_2, \ldots$ before A, with a similar definition for $Awins$. If B is of the same length as A and $\mathbb{P}(Bwins) > \mathbb{P}(Awins)$, we say that B *beats* A. Given A, if there exists a word B that beats A we say that A *can be beaten*. The question as to whether every word A can be beaten is the basis of the game Penney-ante [6].

† The author thanks Pavel Pevzner for his conjecture, Richard Arratia for encouragement, and the referees for helpful suggestions.

When the X_i are uniformly distributed, Guibas and Odlyzko [4] have shown that every word of length at least three can be beaten. Clearly, though, there are non-uniform distributions under which some longer words cannot be beaten. For example, if $A = b^n$ consists of n repetitions of a letter b, then any distribution for which $p(b) > (0.5)^{1/n}$ has $\mathbb{P}(T_A = n) > 1/2$, so A cannot be beaten.

It is conjectured in Pevzner [7] that for any nontrivial probability distribution $\mathbb{P}$ there is a positive integer N such that any word whose length is at least N can be beaten. Pevzner's conjecture happens to be true if $|S| = 2$ or $|S| = 3$, but for higher values of $|S|$ it may be true or false, depending on the relative sizes of the mode and secondary mode of the distribution. If the mode is much larger than the secondary mode, then the conjecture does not hold, but only fails for the word A consisting solely of repetitions of the letter of maximum likelihood.

To state our results, we need notation for the mode and the secondary mode. For the mode, we write $p = \max\{p(s) : s \in S\}$. Let $M = \{s : p(s) = p\}$ be the set of letters attaining the mode. The secondary mode is defined by fixing any $b \in M$ and setting $c \equiv \max\{p(s) : s \in S - \{b\}\}$. Thus, by our assumption that $|S| \geq 2$ and $p(s) > 0$ for all $s \in S$, we have $0 < c \leq p < 1$.

Theorem 1. *Suppose $\mathbb{P}$ is a probability distribution on a set S with $|S| \geq 2$. If $c > p\frac{1-p}{1+p}$, then there is a positive integer $N > 0$ such that every word of length $n \geq N$ can be beaten.*

If $c \leq p\frac{1-p}{1+p}$, then the inequality $c < p$ shows there is a unique element b of maximal likelihood. For each positive n the word b^n is not beaten by any other word of length n. However, there is a positive integer $N > 0$ such that every word of length $n \geq N$ not equal to b^n can be beaten.

The inequality $c > p\frac{1-p}{1+p}$ holds whenever $|S| = 2$ or $|S| = 3$, and probability distributions are easily constructed for which $c \leq p\frac{1-p}{1+p}$ on any space S with $|S| > 3$. The inequality $c > p\frac{1-p}{1+p}$ may be expressed in the following way: If $b \in M$ is assigned probability p, then the 'remaining' probability is $1 - p$, and c must be larger than $\frac{p}{1+p}$ times the 'remaining' probability for the inequality to hold. This interpretation has the attraction that $\frac{p}{1+p}$ is monotonically increasing, as opposed to $p\frac{1-p}{1+p}$.

For a reduced pair of words A and B, and for X_i distributed uniformly, Conway gave an explicit formula for the ratio $\mathbb{P}(B\text{wins})/\mathbb{P}(A\text{wins})$. We say that a pair of words is *reduced* if neither is a subword of the other. (We say that $A = a_1 \ldots a_n$ is a subword of $B = b_1 \ldots b_\eta$ if for some $1 \leq k \leq \eta - n + 1$, we have $a_i = b_{k+i}$ for all $1 \leq i \leq \eta$.) For words of the same length, as in Theorem 1, a pair of words is reduced if and only if they are distinct. A version of Conway's formula applying to X_i having any distribution, stated as Theorem 2, is needed for our results. Li [5] uses martingale methods to prove Theorem 2 and a generating function approach is used in Guibas and Odlyzko [4]; Breen *et al.* [1] use similar generating function methods to determine the expected recurrence time for a set of reduced patterns.

The probability of a nonempty word is denoted throughout this paper by

$$\mathbb{P}(a_1 a_2 \ldots a_n) = p(a_1)p(a_2)\cdots p(a_n),$$

and the probability of the empty word is set to 1. The following notation is needed for Theorem 2:

$$\varepsilon(A,B,k) = \begin{cases} 1 & \text{if } a_k \dots a_{|A|} = b_1 \dots b_{|A|-k+1} \\ 0 & \text{otherwise,} \end{cases}$$

$$(AB) = \sum_{k=1}^{|A|} \mathbb{P}(b_{|A|-k+2} \cdots b_{|B|})\, \varepsilon(A,B,k).$$

Theorem 2. *If A and B are a reduced pair of words, then*

$$\frac{\mathbb{P}(Bwins)}{\mathbb{P}(Awins)} = \frac{\mathbb{P}(B)(AA) - \mathbb{P}(A)(AB)}{\mathbb{P}(A)(BB) - \mathbb{P}(B)(BA)}. \tag{1}$$

To illustrate Theorem 2, let $S = \{a,b\}$, $A = bab$ and $B = bba$, and suppose that the letters have probabilities $p(a) = 1/3$ and $p(b) = 2/3$. We then have $(AA) = 1 + 2/9 = 11/9$, $(AB) = 2/9$, $(BA) = 2/3$, $(BB) = 1$, $\mathbb{P}(A) = \mathbb{P}(B) = 4/27$. In this example, $\mathbb{P}(Bwins)/\mathbb{P}(Awins) = 3$, i.e. B beats A.

Theorem 1 will be proved by first examining the case $A = b^n$ with $b \in M$ in Lemma 1 and then verifying that Pevzner's conjecture holds for all other words in Proposition 1. The condition $c > p\frac{1-p}{1+p}$ is partially explained with the following argument. Suppose $b \in M$ and $a \in M - \{b\}$ with $p(a) = c$. For large n we may assume that first occurrence of the string b^{n-1} is not at the beginning of the sequence $X_1, X_2, \dots$. The probability that the appearance of the word b^{n-1} that determines which of b^n or ab^{n-1} occurs first is preceded by an a is $\frac{c}{1-p}$; the probability that it is succeeded by a b and not preceded by an a is $(1 - \frac{c}{1-p})p$. Setting these probabilities equal to each other produces the equation $c = p\frac{1-p}{1+p}$. As shown in Lemma 1, if $A = b^n$, then choosing $B = ab^{n-1}$ maximizes $\mathbb{P}(Bwins)/\mathbb{P}(Awins)$.

Lemma 1. *Suppose $b \in M$, and $a \in S - \{b\}$ is such that $p(a) = c$. If $c > p\frac{1-p}{1+p}$, then ab^{n-1} beats b^n for large enough n. If $c \le p\frac{1-p}{1+p}$, then for all $n > 0$, the word b^n is not beaten by any word of length n.*

Proof. Let A $= b^n$ and B $= ab^{n-1}$. By direct calculation, we have $(AA) = 1 + p + \cdots + p^{n-1}$, $(AB) = 0$, $(BB) = 1$, and $(BA) = p + \cdots + p^{n-1}$. From Theorem 2,

$$\frac{\mathbb{P}(Bwins)}{\mathbb{P}(Awins)} = \frac{c(1 + p + \dots + p^{n-1})}{p - c(p + \dots + p^{n-1})}. \tag{2}$$

If we let $f(x) = \frac{c(1+x)}{p-cx}$, then $f(x)$ is defined for $x = p + p^2 + \cdots + p^{n-1}$, as the inequality $p + p^2 + \cdots + p^{n-1} < p/(1-p) \le p/c$ shows. As $n \to \infty$, $\mathbb{P}(Bwins)/\mathbb{P}(Awins) \to (\frac{c}{1-p})/(p - \frac{cp}{1-p}) = \frac{c}{p-p(c+p)}$. This expression is greater than 1 (or is infinite) if and only if $c > p\frac{1-p}{1+p}$. Because $f(x)$ is monotone increasing, when $c > p\frac{1-p}{1+p}$ the required N exists and when $c \le p\frac{1-p}{1+p}$ equation (2) is less than 1 for all n. For $A = b^n$, choosing $B = ab^{n-1}$ minimizes

(AB) and (BB) and maximizes (BA) and $\mathbb{P}(B)$; if $B = ab^{n-1}$ cannot beat A, neither can any other choice of B. □

Remark. For any $N_0 > 0$, we can construct a probability distribution with $c > p\frac{1-p}{1+p}$ for which the N in Theorem 1 must satisfy $N \geq N_0$. It follows from equation (2) that $A = b^n$ can be beaten if and only if

$$p < c\left(1 + 2p + 2p^2 + \cdots + 2p^{n-1}\right)$$

or, equivalently,

$$p^{n-1} < 1 - \frac{(p-c)(1-p)}{2pc}. \tag{3}$$

Given any $0 < p < 1$ and $0 < s < 1$, a probability distribution $\mathbb{P}$ may be constucted on a large enough space S with mode p and secondary mode $c = s(1-p)$. Suppose that $s = \frac{3p}{2(1+p)}$. For this choice of s, we have $c > p\frac{1-p}{1+p}$, and for $p > 1/2$ the right-hand side of equation (3) is positive and bounded by $1/2$, so that $N > 1 + \log_p(1/2)$. Letting p approach 1 forces N to become arbitrarily large.

In the rest of this paper, we complete the proof of Theorem 1 by showing that there exists an N such that any word of length at least N, and not of the form $A = b^n$ with $b \in M$, can be beaten. One might suppose from the heuristic stated just before Lemma 1 that for any $b \in M$, $B = ba_1a_2 \ldots a_{n-1}$ will beat $A = a_1a_2 \ldots a_n$ when A is long enough. This last remark is true, except for special cases of A for which it is possible that $\mathbb{P}(Bwins)/\mathbb{P}(Awins) = 1$; for these special cases we need only be more specific in our choice of b to make the heuristic work. Guibas and Oldyzko [4] used a variant of this idea to show that such a word B beats A for X_i distributed uniformly.

It will be useful to write a word A as multiple concatenations of a subword whose length is the basic period of A. The concatenation of two words $A = a_1a_2 \ldots a_n$ and $B = b_1b_2 \ldots b_\eta$ is $AB = a_1a_2 \ldots a_nb_1b_2 \ldots b_\eta$. The basic period α of word $A = a_1a_2 \ldots a_n$ is defined to be the size of the smallest shift of A such that the shifted word overlaps the original word, if such a shift exists, or n if it does not. Any word A may be written as $A = T^mT^*$ where $T = a_1a_2 \ldots a_\alpha$, $T^* = a_1a_2 \ldots a_\beta$ with $1 \leq \beta \leq \alpha$ and m is a non-negative integer; see Guibas and Odlyzko [3]. It follows from the definition of T that for $m \geq 2$, the only shifts of T of size smaller or equal to $(m-1)\alpha$ overlapping T^m are exactly those shifts $j\alpha$ for $j = 1, 2, \ldots, m-1$. This informal discussion is summarized more precisely in the following lemma without proof.

Lemma 2. *For a word $A = a_1 \ldots a_n$, let $\mathscr{A} = \{k \in [1, n-1] : a_1 \ldots a_{n-k} = a_{k+1} \ldots a_n\}$ be the set of self overlapping shifts of A. Let α denote the smallest element in $\mathscr{A}$ if $\mathscr{A} \neq \emptyset$, and let $\alpha = n$ if $\mathscr{A} = \emptyset$. If $T = a_1a_2 \ldots a_\alpha$, then A may be written uniquely as T^mT^* for some integer $m \geq 0$, where $T^* = a_1 \ldots a_\beta$ for some $\beta \in [1, \alpha]$. Furthermore, there is no $k \in [1, \alpha - 1]$ such that $a_1 \ldots a_{\alpha-k} = a_{k+1} \ldots a_\alpha$ and $a_{\alpha-k+1} \ldots a_\alpha = a_1 \ldots a_k$.*

From now on, when $A = T^mT^*$ is written, it is meant that T^mT^* is the unique form of A given by Lemma 2.

Proposition 1. *There exists an $N > 0$ such that any word A that is not of the form $A = b^n$ for some $b \in M$ and whose length is at least N can be beaten.*

Proof. By letting $T^\dagger$ denote $a_1 a_2 \ldots a_{\beta-1}$ when $\alpha > 1$ and the empty word when $\alpha = 1$, for any $b \in M$ we may write $b a_1 a_2 \ldots a_{n-1}$ more compactly as $bT^m T^\dagger$. Introducing the notation $[A] \equiv (AA) - 1$ and letting $q = p(a_\beta)$, we have the following equations:

$$(AA) = 1 + [T^m T^*], \tag{4}$$

$$(AB) = \frac{1}{q}[bT^m T^*], \tag{5}$$

$$(BB) = 1 + [bT^m T^\dagger], \tag{6}$$

$$(BA) = q(1 + [T^m T^\dagger]). \tag{7}$$

Equation (5) follows from observing that concatenating a_β to the end of B produces $bT^m T^*$, and that the shifts involved with the calculation of $(A\ bT^m T^*)$ correspond exactly with those in the calculation of $[bT^m T^*]$. Equation (7) is obtained in a similar manner. Inserting equations (4–7) into equation (1) and dividing numerator and denominator by $\mathbb{P}(A) = \mathbb{P}(T)^m \mathbb{P}(T)$ gives us the identity

$$\frac{\mathbb{P}(Bwins)}{\mathbb{P}(Awins)} = \frac{p(1 + [T^m T^*]) - [bT^m T^*]}{q(1 + [bT^m T^\dagger]) - pq(1 + [T^m T^\dagger])}.$$

To avoid writing complicated fractions, we define ρ to be

$$\rho \equiv q[bT^m T^\dagger] + [bT^m T^*] - p[T^m T^*] - pq[T^m T^\dagger] + q - p - pq,$$

so that

$$\rho < 0 \quad \Rightarrow \quad \mathbb{P}(Bwins)/\mathbb{P}(Awins) > 1.$$

Different arguments are used for bounding ρ when $m = 0$, $m = 1$ and $m \geq 2$.

Words A for which $m = 0$ have no self-overlap, and therefore $[T^*] = 0$ and $[bT^*] \leq \mathbb{P}(T^*)$. The possible self-overlapping shifts of $bT^\dagger$ are restricted by the self-overlapping shifts of $T^\dagger$, leading to the bound $[bT^\dagger] \leq [T^\dagger] + \mathbb{P}(T^\dagger)$. If the first shift of $T^\dagger$ overlaps itself, then $T^\dagger$ must be of the form $T^\dagger = a_1^{n-1}$ for some $a_1 \in S$, in which case, $A = a_1^{n-1} a_2$ with $a_1 \neq a_2$. One may check by direct calculation with equation (1) for $n \geq 3$ that $B = a_2 a_1^{n-1}$ beats $A = a_1^{n-1} a_2$ whenever $a_2 \in M$; that $B = a_1^n$ beats A whenever $a_1 \in M$ and $a_2 \notin M$; and that $B = b a_1^{n-1}$ beats A for any $b \in M$ whenever $a_1 \notin M$ and $a_2 \notin M$. Assuming that the first shift of $T^\dagger$ does not overlap itself gives us the bound $[T^\dagger] < p^2/(1-p)$. Choosing N large enough so that $p^{N-1} \leq (p - p^2)/2$, for all words A with $M = 0$ with length at least N, we have

$$\begin{aligned} \rho &\leq q[bT^\dagger] + [bT^*] - pq[T^\dagger] - pq \\ &\leq q(1-p)[T^\dagger] + q\mathbb{P}(T^\dagger) + \mathbb{P}(T^*) - pq \\ &< qp^2 + q(p - p^2) - pq \\ &= 0. \end{aligned}$$

In the argument bounding ρ for words with $m = 1$, we may assume that the first

self-overlap k of $bTT^\dagger$ satisfies $k \geq 3$, for the reason that if A is of the form $A = abab\ldots$, then A is beaten by $B = baba\ldots$ if $a \notin M$ and by $B = aabab\ldots$ if $a \in M$. Since T is assumed not to be of the form b^n with $b \in M$, we know that there exists a letter a' in T for which $p(a') = r \leq 1 - p$, and that the first shift of $bTT^\dagger$ that overlaps itself must 'push' the last occurrence of a' past the last letter of $bTT^\dagger$. We will make use of these definitions of a' and r in the proof with $m \geq 2$ as well.

The following bound holds for words of length at least N, with $m = 1$ supposing that N is large enough so that $p^{N/2-1} \leq p^2 - p^3$:

$$\begin{aligned}
\rho &\leq q[bTT^\dagger] + [bTT^*] + q(1-p) - p \\
&\leq p[bTT^\dagger] + [bTT^*] + p(1-p) - p \\
&< p^3 r/(1-p) + p^{|A|/2-1} r/(1-p) - p^2 \\
&\leq p^3 + (p^2 - p^3) - p^2 \\
&= 0.
\end{aligned}$$

For words with $m \geq 2$, we may assume that $\mathbb{P}(T) < pq$; else we have the special case $A = abab\ldots$ with $a \in M$. Because S is finite, there exists some $K > 0$ such that $pq - \mathbb{P}(T) \leq K < pq$ for all T such that $\mathbb{P}(T) < pq$. Letting $\kappa = |T^{m-1}|$, we have the inequality

$$\kappa = (m-1)|T| = (1 - m^{-1})|T^m| \geq \frac{1}{4}|A|.$$

Suppose that N is large enough so that $2p^{N/4} \leq K$. For all words A with $m \geq 2$ whose length is at least N,

$$\begin{aligned}
\rho &\leq \Big(q[bT^m T^\dagger] - p[T^m T^*]\Big) + \Big([bT^m T^*] - pq[T^m T^\dagger] - pq\Big) \\
&\leq \mathbb{P}(T)^{m-1}\Big(q[bTT^\dagger] - p[TT^*]\Big) + \Big(\mathbb{P}(T) - pq\Big) \\
&\quad + \mathbb{P}(T)^{m-1}\Big([bTT^*] - pq[TT^\dagger]\Big) - pq\mathbb{P}(T)^{m-1} \\
&\leq q\mathbb{P}(T)^{m-1}[bTT^\dagger] + \mathbb{P}(T)^{m-1}[bTT^*] - K. \\
&< 2(rp^{\kappa-1})p/(1-p) - K \\
&\leq 2p^\kappa - K \\
&\leq 0.
\end{aligned}$$

This completes the proof of Proposition 1. □

References

[1] Breen, S., Waterman, M.S. and Zhang, N. (1985) Renewal Theory for Several Patterns. *J. Appl. Prob.* **22** 228–234.

[2] Chen, R. (1989) A Circular Property of the Occurrence of Sequence Patterns in the Fair Coin-Tossing Process. *Adv. Appl. Prob.* **21** 938–940.

[3] Guibas, L.J. and Odlyzko, A.M. (1981) Periods in Strings. *J. Comb. Theory A* **30** 19–42.

[4] Guibas, L.J. and Odlyzko, A.M. (1981) String Overlaps, Pattern Matching and Nontransitive Games, *J. Comb. Theory A* **30** 183–208.

[5] Li, S.-Y. R. (1980) A Martingale Approach to the Study of Occurrence of Sequence Patterns in Repeated Experiments. *Annal. Prob.* **8** 1171–1176.
[6] Penney, W. (1969) Problem: Penney-ante. *J. Rec. Math.* **2** 241.
[7] Pevzner, P. (1987) The Best Bet for Simpletons. *Kvantl* **5** 4–15.